

Simplicial Complex Entropy

Stefan Dantchev
Durham University
School of Eng. & Comp. Sciences
Durham, DH1 3LE, UK
s.s.dantchev@durham.ac.uk

Ioannis Ivrissimtzis
Durham University
School of Eng. & Comp. Sciences
Durham, DH1 3LE, UK
ioannis.ivrissimtzis@durham.ac.uk

Abstract

We propose an entropy function for simplicial complices. Its value gives the expected cost of the optimal encoding of sequences of vertices of the complex, when any two vertices belonging to the same simplex are indistinguishable. We show that the proposed entropy function can be computed efficiently. By computing the entropy of several complices consisting of hundreds of simplices, we show that the proposed entropy function can be used in the analysis of the large sequences of simplicial complices that often appear in computational topology applications.

1 Introduction

In several fields of visual computing, such as computer vision, CAD and graphics, many applications require the processing of an input in the form of a set of unorganized points, that is, a finite subset of a metric space, typically \mathbf{R}^2 or \mathbf{R}^3 . Often, the first step in the processing pipeline is the construction of a simplicial complex, or a series of simplicial complices capturing spatial relations of the input points. Such geometrically constructed simplicial complices commonly used in practice include the *Vietoris-Rips* and *Čech* complices, see for example [1], the *alpha shapes* [2] and the witness complices [3, 4].

The two simplest constructions, giving the Vietoris-Rips and the Čech complices, emerged from studies in the field of algebraic topology. In the Vietoris-Rips construction, we connect two points with an edge if their distance is less than a fixed ε and the simplices of the complex are the cliques the resulting graph. In the Čech construction, the simplices are the sets of vertices that lie inside a bounding sphere of radius ε .

Notice that the complices constructed this way, apart from the input point set which gives their vertex set, also depend on the parameter ε . In applications where the goal is to extract topological information related to input point set, it is quite common to consider sequences of complices corresponding to different values of ε and study the evolution of their topological properties as ε varies [5, 6]. Such investigations led to the development of the notion of *persistence*, in the form for example of persistent homology, as one of the main concepts in the field of computational topology [7, 8]. Indicative of the

need for computational efficiency, persistent homology calculations based on millions of distinct complices from the same input point set are now common and thus, the efficient computations of such series of complices is an active research area [1, 9].

In this paper, our aim is to use information theoretic tools to study sequences of geometrically constructed complices corresponding to different values of ε . In particular, we define an entropy function on simplicial complices; we show that it can be computed efficiently; and demonstrate that it can be used to find critical values of ε . Here, the value of ε is seen as a measure of spatial resolution and thus, we interpret the simplices of the geometrically constructed complices as sets of indistinguishable points.

The setting of our problem is very similar to one that gave rise to the concept of *graph entropy* [10] and *hypergraph entropy* [11]. There, a graph or a hypergraph describe indistinguishability relations between vertices and the sets of indistinguishable vertices are derived as the *independent sets* of the graph or hypergraph. In contrast, in our approach, the sets of indistinguishable vertices are readily given as the simplices of the complex. In the next section, immediately after introducing the proposed simplicial complex entropy, we discuss in more detail its relation to graph entropy.

2 Simplicial complex entropy

Let $V = \{v_1, \dots, v_n\}$ be a point set consisting of n vertices. An abstract simplicial complex C over V is given by its maximal simplices C_1, \dots, C_m . These are nonempty subsets of V whose union is the entire V and none of them is a subset of another.

We are also given a probability distribution P over V , i.e. non-negative numbers p_1, \dots, p_n and such that $\sum_{j=1}^n p_j = 1$. Assuming that all points that belong to the same simplex C_i for some i , $1 \leq i \leq m$ are indistinguishable, we define the *simplicial complex entropy* as

$$H(C, P) = \min \sum_{j=1}^n p_j \log \frac{1}{\sum_{i \in \text{Simpl}(j)} q_i} \quad \text{s.t.} \quad (1)$$

$$\sum_{i=1}^m q_i = 1$$

$$q_i \geq 0 \quad 1 \leq i \leq m.$$

where $\text{Simpl}(j)$ denotes the set of simplices containing vertex p_j .

The above simplicial complex entropy is similar to the graph entropy, defined over a graph G with a probability distribution P on its vertices, given by

$$\min \sum_{j=1}^n p_j \log \frac{1}{a_i} \quad (2)$$

where the minimum is taken over all convex combinations of characteristics vectors \mathbf{a} of the independent sets of G , with a_i denoting the i -th coordinate of such vectors.

In its information theoretic interpretation, the graph entropy gives the expected number of bits per symbol required in an optimal encoding of the information coming from a source emitting vertices of G under the probability distribution P , assuming that any two vertices are indistinguishable iff they are not connected with an edge [12]. In other words, the independent sets of G are the sets of mutually indistinguishable vertices. Similarly, the information theoretic interpretation of the proposed simplicial complex entropy is that of the expected bits per vertex required in an optimal encoding of the information coming from the same source, under the assumption that the sets of mutually indistinguishable vertices are exactly the simplices.

The proposed simplicial complex entropy can be seen as a simplification of the graph entropy, which however is at least as general. Indeed, on a graph G we can define a simplicial complex C on the same vertex set as G and its simplices being the independent sets of G . Then, the graph entropy of G is the simplicial complex entropy of C . On the other hand, given a simplicial complex C it is not immediately obvious how one can construct a graph G such that the simplicial complex entropy of C is the graph entropy of G .

In an abstract context, the proposed simplification might seem quite arbitrary: instead of deriving the sets of indistinguishable vertices from the connectivity of a graph, we consider them given in the form of simplices. However, in the context of geometrically constructed simplicial complexes embedded in a metric space, the simplices are the natural choice of sets of indistinguishable points for a given spatial resolution ε and there is no need, or indeed an obvious way, to model the property of indistinguishability in terms of graph connectivity. One notable exception to this is the special case of Vietoris-Rips complexes which we discuss next, aiming at further highlighting differences and similarities between simplicial entropy and graph entropy.

2.1 Example: Vietoris-Rips simplicial complex entropy

In the case of Vietoris Rips complexes, there is a straightforward interpretation of the simplicial complex entropy as graph entropy. Indeed, assume a probability distribution P on a set of vertices V embedded in a metric space, and assume that two vertices, are indistinguishable if their distance is less than ε . The graph G with its edges connecting pairs of distinguishable vertices is the complement of the underlying graph of the Vietoris-Rips complex constructed on V for the same ε .

It is easy to see that the independent sets of G are exactly the simplices of the Vietoris-Rips complex and thus, the graph entropy of G is the simplicial complex entropy of the Vietoris-Rips complex. Indeed, if there are no edges connecting points of a subset of V , it means that all distances between these points are less than ε , therefore they form a simplex of the Vietoris-Rips complex.

The simplicial entropy of the Vietoris-Rips complexes has a straightforward graph entropy interpretation because Vietoris-Rips complexes are completely defined by their underlying graph. Indeed, their simplices are the cliques of the underlying graph. However, this is not generally the case for geometrically constructed complexes, with the Čech complex being a notable counterexample.

Indeed, consider as V the three vertices of an equilateral triangle of edglength 1, embedded in \mathbf{R}^2 . Any pair of vertices corresponds to an edge of the triangle and has a minimum enclosing sphere of radius $1/2$. The V itself has a minimum enclosing sphere of radius $\sqrt{3}/3$. Thus, for any $1/2 \leq \varepsilon \leq \sqrt{3}/3$ all three edges of the triangle are simplices of the Čech complex, i.e. pair-wise indistinguishable, but the triangle itself is not a simplex of the Čech complex.

3 Properties of simplicial complex entropy

Solving the entropy minimisation turns out to be computationally tractable. Let us denote

$$S_j(q) \stackrel{\text{def}}{=} \sum_{i \in \text{Simpl}(j)} q_i$$

and rewrite (1) as a maximisation problem with an objective function

$$f(q) \stackrel{\text{def}}{=} \sum_{j=1}^n p_j \log S_j(q). \quad (3)$$

We can immediately prove the following

Proposition 1. *The objective function (3) is concave. The sums $S_j(q)$ are unique (i.e. the same) for all points q where the maximum is attained, while the set of all maxima is a polyhedron.*

Proof. Let q' and q'' be two different feasible points. Clearly, the point $q \stackrel{\text{def}}{=} 1/2(q' + q'')$ is also feasible and $S_j(q) = 1/2(S_j(q') + S_j(q''))$ for $1 \leq j \leq n$. We then have

$$\log S_j(q) \geq 1/2(\log S_j(q') + \log S_j(q'')), \quad (4)$$

which proves the concavity of the objective function.

Imagine now that q' and q'' are two (different) optimal points (with $f(q') = f(q'')$) and moreover there is a j , $1 \leq j \leq n$ such that $S_j(q') \neq S_j(q'')$. For that particular, (4) is a strict inequality and after summing up all inequalities, we get

$$f(q) > 1/2(f(q') + f(q'')),$$

which contradicts the optimality of both q' and q'' . Thus, the sums $\log S_j(q)$ are unique over all optimal points q .

Finally, if we denote these sums (at an optimum) by s_j , $1 \leq j \leq n$, we notice that the set of all optimal points q is fully described by the following linear system:

$$\begin{aligned} \sum_{i \in \text{Simpl}(j)} q_i &= s_j & 1 \leq j \leq n \\ \sum_{i=1}^m q_i &= 1 \\ q_i &\geq 0 & 1 \leq i \leq m. \end{aligned}$$

□

Another useful characterisation of an optimal point q is given by

Proposition 2. *Any optimal point q satisfies the following “polynomial complementarity” system:*

$$\begin{aligned} \sum_{j \in Pts(i)} \frac{p_j}{S_j(q)} & \begin{cases} = 1 & \text{if } q_i > 0 \\ \leq 1 & \text{if } q_i = 0 \end{cases} & 1 \leq i \leq m \\ \sum_{i=1}^m q_i &= 1 \\ q_i &\geq 0 & 1 \leq i \leq m \end{aligned}$$

Proof. The gradient of the objective function, $\nabla f(q)$ is

$$\left(\sum_{j \in Pts(1)} p_j / S_j(q), \dots, \sum_{j \in Pts(m)} p_j / S_j(q) \right)^T.$$

We start with Karush–Kuhn–Tucker conditions (for the maximisation problem) that an optimal point q should satisfy:

$$\sum_{j \in Pts(i)} \frac{p_j}{S_j(q)} = \lambda - \mu_i \quad 1 \leq i \leq m \quad (5)$$

$$\sum_{i=1}^m q_i = 1 \quad (6)$$

$$q_i, \mu_i \geq 0 \quad q_i \mu_i = 0 \quad 1 \leq i \leq m \quad (7)$$

for some λ and μ_i , $1 \leq i \leq m$.

We first expand the inner product

$$\langle q, \nabla f(q) \rangle = \sum_{i=1}^m q_i \sum_{j \in Pts(i)} \frac{p_j}{S_j(q)} = \quad (8)$$

$$= \sum_{j=1}^n \frac{p_j}{S_j(q)} \sum_{i \in Simpl(j)} q_i = \sum_{j=1}^n p_j = 1. \quad (9)$$

On the other hand, from 5, 6 and 7, we get

$$\sum_{i=1}^m q_i \sum_{j \in Pts(i)} \frac{p_j}{S_j(q)} = \sum_{i=1}^m q_i (\lambda - \mu_i) = \quad (10)$$

$$= \lambda \sum_{i=1}^m q_i - \sum_{i=1}^m q_i \mu_i = \lambda, \quad (11)$$

and thus $\lambda = 1$.

□

3.1 Defining the error

The indistinguishability between points, as described by the simplicial complex, results into an error of a complex can be understood in terms of encoding and decoding points as follows.

The encoder gets point j ($1 \leq j \leq n$), which is produced by a memoryless random source under distribution p . We will describe two encoding strategies, one randomised, which is the one we implemented, and an adversarial which should give higher error rates.

3.1.1 Randomised encoder

The randomised encoder produces one of the cells that contains j , under the distribution $q_i/S_j(q)$ for all $i \in \text{Simpl}(j)$. The overall probability of seeing cell i as a result is

$$\sum_{j \in \text{Pts}(i)} p_j \frac{q_i}{S_j(q)} = q_i \sum_{j \in \text{Pts}(i)} \frac{p_j}{S_j(q)} = q_i (1 - \mu_i) = q_i \quad (12)$$

(where μ_i is as in the proof of proposition 2 above and taking into account that $\lambda = 1$) as expected.

The decoder sees a cell i and its best guess (as to which point actually produced it) is the one that has the biggest probability (according to the distribution p). Thus the total gain is

$$\text{err} = \sum_{i=1}^m q_i \frac{\max_{j \in \text{Pts}(i)} p_j}{\sum_{j \in \text{Pts}(i)} p_j}.$$

3.1.2 The adversarial encoder

We can think of this encoding strategy as a game between the encoder and the decoder, in which whenever the decoder sees a simplex i , he responds with a guess (point) $j \in \text{Simpl}(i)$ according to probabilities r_{ij} , $r_{ij} \geq 0$ and such that

$$\sum_{j \in \text{Pts}(i)} r_{ij} = 1 \quad \text{for every } 1 \leq i \leq m.$$

These probabilities are known to the encoder, so if the source produced a point j , the encoder minimises the gain of the decoder by picking a cell i that is $\arg \min_{j \in \text{Simpl}(i)} r_{ij}$. In turn, the decoder tries to maximise their total expected gain as

$$\begin{aligned} \overline{\text{err}} = \max \sum_{j=1}^n p_j r_j \quad \text{s.t.} \\ r_{ij} \geq r_j \quad & 1 \leq j \leq n \text{ and } i \in \text{Simpl}(j) \\ \sum_{j \in \text{Pts}(i)} r_{ij} = 1 \quad & 1 \leq i \leq m \\ r_{ij} \geq 0 \quad & 1 \leq j \leq n \text{ and } i \in \text{Simpl}(j) \end{aligned}$$

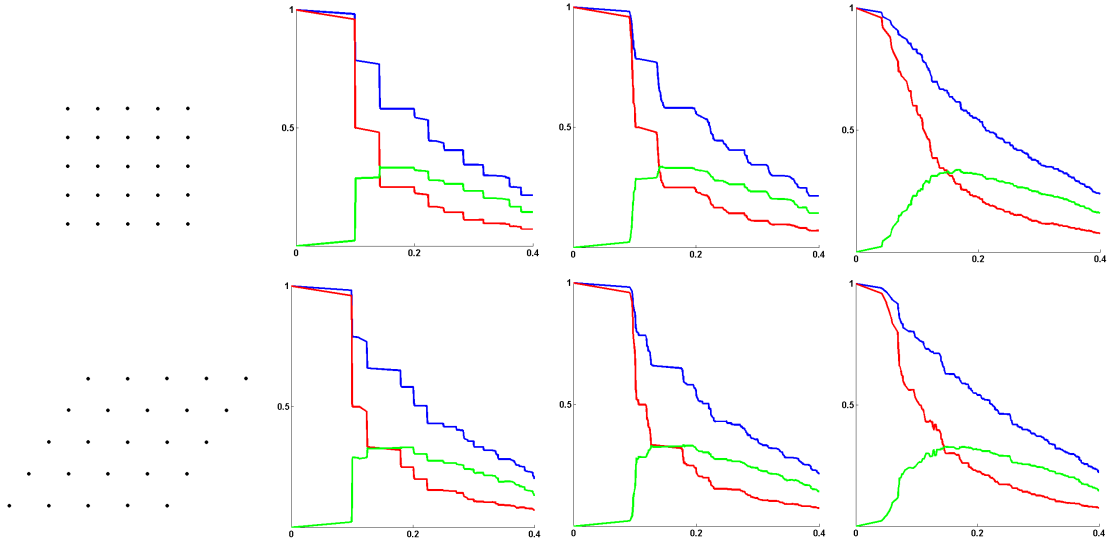


Figure 1: The y -axis represents the normalised entropy (blue curve), the accuracy rate (red curve) and their difference (green curve). The x -axis represents the parameter ε (radius of the minimal enclosing sphere of a simplex) in the construction of the Čech complex. **Top:** The input point set is the 5×5 block of vertices of a square grid of edglength 0.2 shown in the left. From left to right, uniform random noise $\pm 0.5\%$, $\pm 5\%$ and $\pm 50\%$ of the edglength was added. The figures represent entropy and error computations on all possible Čech complices for that range of ε , that is, 768, 746 and 685 distinct complices, respectively. **Bottom:** As per the top, but for triangular grid points. The figures correspond to 725, 694 and 672 distinct complices, respectively.

4 Examples

The computation of the simplicial complex entropy and the error was implemented in Matlab. Apart from some code for input output operations and simplicial complex representation, *fmincon* and *linprog* were directly used to compute the entropy and the error, respectively.

In all examples, we report the *normalised entropy*, that is, the simplicial complex entropy $H(C, P)$ divided by the entropy of the vertex set V under the same probability distribution P . Instead of the randomised encoder error err in Eq. 12, we report the value $1 - err$, which can be seen as the decoding accuracy rate and correlates nicely with the normalised entropy. The difference between these two values is also reported.

In a first example, Figure 1 (Top) shows the values of these two functions on Čech complices constructed from vertex sets that are nodes of square grid of edglength 0.2 with some added noise. Figure 1 (Bottom) shows a similar example with the vertices originally being nodes of a triangular grid. In all cases, the probability distribution P on the vertex set is uniform.

In the case of a square grid without any added noise, as the values of the parameter

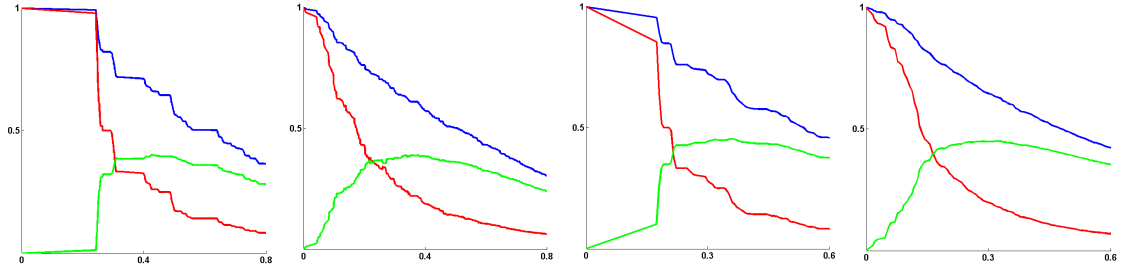


Figure 2: The axes and the colour of the curves are as per Figure 1. **Two left figures:** The input point set of size 50 is a computational solution to the Thomson problem with uniform noise of ± 0.01 units added on each coordinate. In the right figure the input is an area uniform spherical random sample of the same size. The figures represent entropy and error computations from 2523 and 2661 distinct Čech complices, respectively. **Two right figures:** As per the top, with point sets of size 100. Due to the very large number of distinct Čech complices, each figure represents 100 Čech complices, corresponding to a uniform sample of values of ε in $[0, 0.6]$.

As ε of the Čech complex construction parameter increase, they reach the first critical value at $\varepsilon = 0.1$, when edges, i.e. simplices of degree 2, are formed. The next critical value is $\varepsilon \simeq 0.141$, where the simplices of degree 4 are formed, and the next critical value is $\varepsilon = 0.2$ when simplices of degree 5 are formed. Similarly, the first critical values in the case of points from a triangular grid are $\varepsilon = 0.1$, when simplices of degree 2 are formed and $\varepsilon \simeq 0.115$ when simplices of degree 3 are formed.

These critical values are shown Figure 1 as sudden drops in the entropy of the Čech complices constructed on the less noisy data sets. We also notice simultaneous drops of the accuracy rates since they, as expected, correlate well with entropy. As the level of noise increases the critical points become less visible on either of these two curves. However, their difference, shown in green, seems to be more robust against noise, and moreover, seems to peak at a favorable place. That is, it peaks in values of ε that would neither return a large number of non-connected components nor heavily overlapping simplices.

In the second example, the input set is a sample from the unit sphere in \mathbf{R}^3 . Figure 2 (left) shows results from regular samples of size 50 (top) and 100 (bottom), computed in [13] as solutions to the Thomson problem, with added uniform noise of ± 0.01 units. In [13], the minimum distances between a point and its nearest neighbour are ~ 0.5 and \sim , respectively, and correspond to the steep entropy decreases and the half of these values when the first edges of the Čech complices are formed. Figure 2 (right) shows results from random, area uniform samples of size 50 (top) and 100 (bottom). While the input is much less regular than at the left hand side of the figure, the peaks of the two green curves align well.

In a third example, we solve the optimisation problem for the computation of the entropy on triangle meshes and show the values of q , as in Eq. 2, color-mapped on the

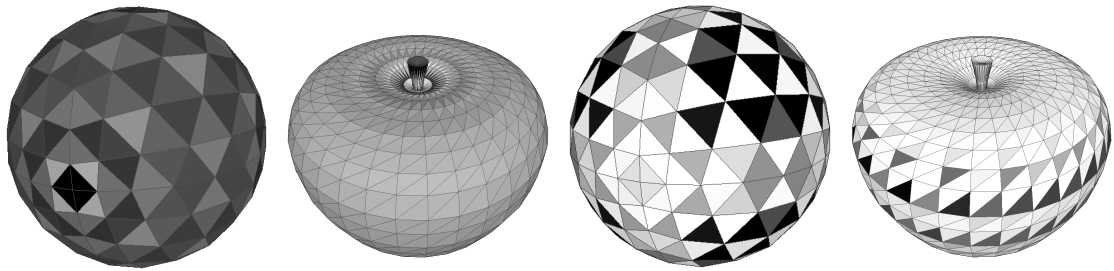


Figure 3: **Two left figures:** The values q_i in Eq. 1 are color-mapped on the mesh triangles. Darker colors correspond to higher values. **Two right figures:** The probability distribution P on the vertices corresponds to the absolute values of the discrete Gaussian curvature of the vertices. The two meshes consist of 512 and 1704 triangle, respectively.

mesh triangles. In Figure 3 (left), the probability distribution on the mesh vertices is uniform, as it was in all previous examples. On the right hand side of the figure, the probability distribution follows the absolute value of the discrete Gaussian curvature of the vertices.

5 Conclusion

We presented an entropy function for simplicial complexes which can be seen as a simplification and generalisation of the graph entropy since all the maximal sets of indistinguishable points are exactly the maximal simplices of the complex and do not have to be computed as the independent sets of the graphs, which, generally, are difficult to characterise. We show that this simplification makes the simplicial complex entropy an function that can be efficiently computed.

Even though the entropy is defined on abstract simplicial complexes, which are purely topological structures, in the examples we show that it can be relevant to geometric applications. For example, by computing the entropy of geometrically constructed simplicial complexes, such as the Čech complexes, or by using geometric properties of an embedded complex, such as a discrete curvature computed on the vertices to obtain a probability distribution on them.

In the future we would like to study in more detail the function given as the difference between normalised entropy and the decoding accuracy rates, which seems to be a robust to noise descriptor of an appropriate level of geometric detail defined by the variable ε of the Čech complex. We would also like to study the relationship between the error corresponding to a randomised encoder we used here and the error corresponding to an adversarial encoder as discussed at the end of Section 3.1.

References

- [1] A. Zomorodian, “Fast construction of the vietoris-rips complex,” *Computers & Graphics*, 2010.
- [2] H. Edelsbrunner, “The union of balls and its dual shape,” *Discrete and Computational Geometry*, vol. 13, no. 1, pp. 415–440, 1995.
- [3] V. de Silva and G. Carlsson, “Topological estimation using witness complexes,” in *Eurographics Symposium on Point-Based Graphics*, M. Alexa and S. Rusinkiewicz, Eds., 2004.
- [4] L. J. Guibas and S. Y. Oudot, “Reconstruction using witness complexes,” in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, ser. SODA '07. Philadelphia, PA, USA: SIAM, 2007, pp. 1076–1085.
- [5] A. Zomorodian and G. Carlsson, “Computing persistent homology,” *Discrete Comput. Geom.*, vol. 33, no. 2, pp. 249–274, 2005.
- [6] M. Vejdemo-Johansson, “Interleaved computation for persistent homology,” *CoRR*, vol. abs/1105.6305, 2011.
- [7] H. Edelsbrunner, D. Letscher, and A. Zomorodian, “Topological persistence and simplification,” in *FOCS '00*. IEEE, 2000, p. 454.
- [8] V. de Silva and R. Ghrist, “Coverage in sensor networks via persistent homology,” *Algebraic and Geometric Topology*, vol. 7, pp. 339–358, 2007.
- [9] S. Dantchev and I. Ivriissimtzis, “Efficient construction of the čech complex,” *Computers & Graphics*, vol. 36, no. 6, pp. 708–713, 2012.
- [10] J. Körner, “Coding of an information source having ambiguous alphabet and the entropy of graphs,” in *6th Prague conference on information theory*, 1973, pp. 411–425.
- [11] J. Korner and K. Marton, “New bounds for perfect hashing via information theory,” *European Journal of Combinatorics*, vol. 9, no. 6, pp. 523–530, 1988.
- [12] G. Simonyi, “Graph entropy: a survey,” *Combinatorial Optimization*, vol. 20, pp. 399–441, 1995.
- [13] D. J. Wales and S. Ulker, “Structure and dynamics of spherical crystals characterized for the thomson problem,” *Physical Review B*, vol. 74, no. 21, p. 212101, 2006.